Probabilistic framework for solving Visual Dialog

Badri N. Patro[1] , Anupriy[2], Vinay P. Namboodiri[2]

[1] *Department of Electrical Engineering, Indian Institute of Technology Kanpur, India*
[2] *Department of Computer Science and Engineering, Indian Institute of Technology Kanpur, India*

Abstract

In this paper, we propose a probabilistic framework for solving the task of 'Visual Dialog'. Solving this task requires reasoning and understanding of visual modality, language modality, and common sense knowledge to answer. Various architectures have been proposed to solve this task by variants of multi-modal deep learning techniques that combine visual and language representations. However, we believe that it is crucial to understand and analyze the sources of uncertainty for solving this task. Our approach allows for estimating uncertainty and also aids a diverse generation of answers. The proposed approach is obtained through a probabilistic representation module that provides us with representations for image, question and conversation history, a module that ensures that diverse latent representations for candidate answers are obtained given the probabilistic representations and an uncertainty representation module that chooses the appropriate answer that minimizes uncertainty. We thoroughly evaluate the model with a detailed ablation analysis, comparison with state of the art and visualization of the uncertainty that aids in the understanding of the method. Using the proposed probabilistic framework, we thus obtain an improved visual dialog system that is also more explainable.

Keywords: CNN, LSTM, Uncertainty, Aleatoric uncertainty, Epistemic Uncertainty Vision and Language, Visual Dialog, VQA, Answer Generation, Question Generation, Bayesian Deep Learning.

Email address: (badri,anupriy,vinaypn)@iitk.ac.in (Badri N. Patro[1] , Anupriy[2], Vinay P. Namboodiri[2])

1. Introduction

Deep learning has aided significant progress in solving various computer vision tasks such as object classification [1, 2] and object detection [3, 4]. The solution of more semantic tasks such as visual question answering [5, 6] and image captioning [7, 8] has also seen progress lately. A challenging problem that extends these is that of maintaining a dialog with a user [9, 10] In this case, a system is required to maintain context concerning the history of the conversation while answering a question and this is be more challenging. A specific task in the visual context is that of the 'Visual Dialog' task [10]. The aim here is that given an image, we need to train an agent to maintain a dialog. The motivation for this emerges from an interest in developing associative technologies for visually impaired persons or chat-bot based dialog agents. Several methods have been proposed for solving the task, such as using various discriminative and generative encoder-decoder frameworks that aim to solve the task of generating dialog [10, 9]. In this paper we aim to extend the previous approaches by formulating a probabilistic approach towards solving this task. This approach is illustrated in figure 1. Through our approach we can obtain a principled model that we can train end-to-end while being able to have uncertainty estimates and the ability to evaluate and explain the model. Such an ability to explain the model is crucial, especially, when we consider that the method could be used by visually impaired people. At any point in the method, the model can be probed to ensure that it is certain about the answers that it generates and more importantly any failure of the method can be addressed by explaining the precise reason for failure as shown in 2. However, as the task is challenging it is important to have an insight into obtaining estimates regarding the uncertainty of the model. This would aid in knowing when the method is confident about its prediction. In this paper, we consider the task of understanding the uncertainty while solving the 'Visual Dialog' task, as shown in 2. This proposed method addresses the limitations of the previous approaches as the previous approaches do not have the ability to obtain uncertainty estimates or obtain diverse conversations.

Our method consists of the following parts :

- Probabilistic Representation Module: Through this module, we obtain proba-

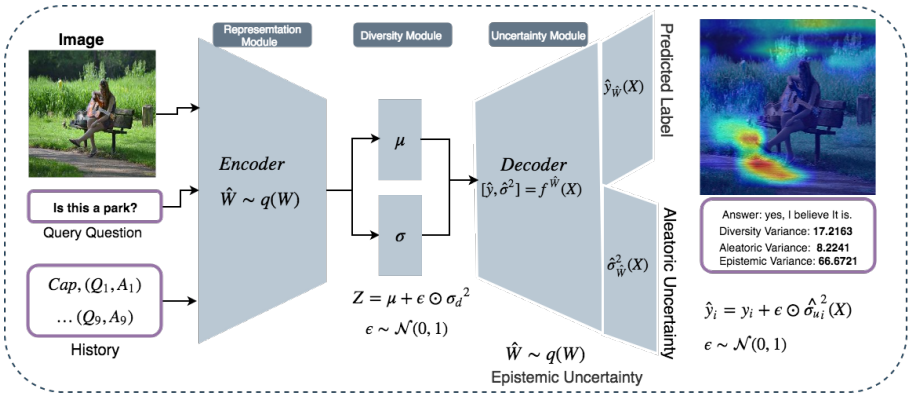

Figure 1: Proposed Probabilistic Diversity and Uncertainty Network (PDUN) consists of three parts, viz. a) Probabilistic Representation Module encodes image feature with a question and history feature in an attentive manner. b) Diversity module captures the diversity, and diverse answer is generated using Variational Auto-Encoder. c) Uncertainty module predicts uncertainty of the network.

bilistic representations for image, question, and history of the conversation using Bayesian CNN and Bayesian RNN modules.

- Diverse Latent Answer Generation Module: In this module, we use a variational autoencoder based latent representation that allows us to obtain latent representations from which we can sample answers.

- Uncertainty Representation Module: In this module, we propose a Reverse Uncertainty based Attention Map (RUAM) method by using Bayesian deep learning methods that allows us to minimize data uncertainty and model uncertainty.

To provide an overview of the technical contributions we make, the main idea is to consider incorporating a Gaussian prior for generating samples of answers. We minimize the KL divergence between the prior and the posterior distribution. The other contribution is to explicitly incorporate a loss to ensure that the correlation between different samples is minimal. We further use these losses along with a loss to minimize the uncertainty. A similar loss has been considered in another context by Patro *et al.* [11]. In this work we are interested in a principled framework for minimizing uncertainty by sampling and generating diverse answers. Moreover, its use has not been considered for the problem of visual dialog. We evaluate each of the contributions

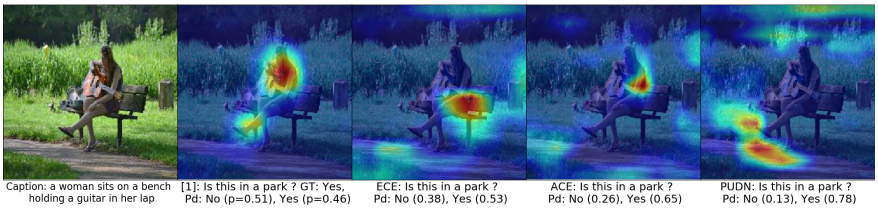

Figure 2: Results were showing the certainty of the correct class increases from baseline[9] to our proposed uncertainty model (PDUN). In this figure, we show the top 2 class confidence score of the question, "Is this a park?". In the baseline model focus on woman, guitar and chair and predicts "NO," which is confused with the correct prediction of the answer, whether it is a park or not. PDUN model minimizes the uncertainty and predicts the correct answer "Yes" with a high confidence score.

in our work. The technical details mentioned here are discussed further in detail in the following sections.

Deep models are usually not interpretable. They are more like black-box models. Due to the lack of transparency, it is difficult to trust the model. We propose a probabilistic method to estimate the uncertainty for the problem of maintaining a dialog with respect to an image, termed the 'Visual dialog' task. In this model, we use gradient certainty based attention model that also improves confidence in the prediction. Using this probabilistic model we do have an improvement in the state of the art results. Moreover, we gain in terms of interpretability by having uncertainty estimates and are also obtain diverse predictions. We have evaluated our method based on the standard matrices as mentioned the visual dialog dataset paper[10]. We observe that , we obtain an improvement in terms of @R10 score around 10% from the baseline 'Late Fusion'[10] model & 3.5% from the State of the art (NMN [12]) method. We also obtain an improvement in terms of NDGC score 9% from base model & 0.5% from State of the art (SOTA) model and in term of MRR, 7% from the base model & 1% from SOTA (NMN [12]) model using our proposed method.

To summarize, the main contributions of this paper are as follows :

1. We provide a module to obtain a probabilistic representation for image, question and conversation history that are obtained as input.
2. The probabilistic representations are used to *generate* diverse latent representations for candidate answers

3. We propose a method for obtaining reverse uncertainty based on attention maps (RUAM). These allow us to select an appropriate answer that minimizes uncertainty.
4. We provide extensive analysis and comparison of our framework with previous methods and evaluate the various ablations of the method. Our proposed framework improves recall@10 score by 3.5%, mean reciprocal rank (MRR) by 1% and NDGC score by 1%. Moreover, we also provide estimation & visualization of the uncertainty of the output.

1.1. Problem Statement: Visual Dialog Task

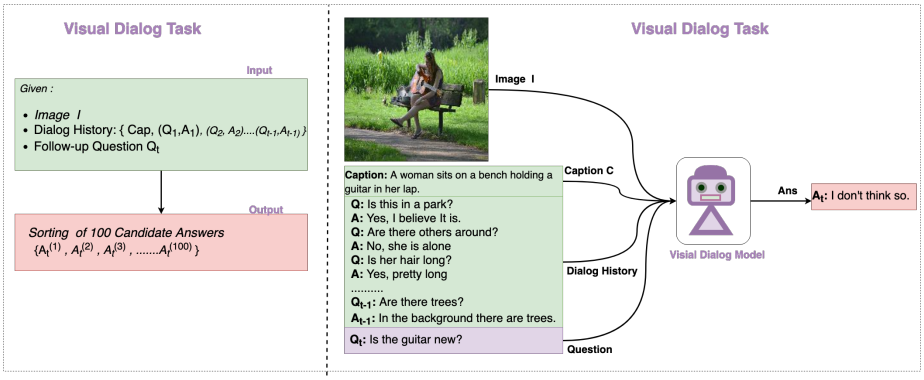

Figure 3: This figure explains the Visual Dialog Task. Left side of the figure explains about, given image I, Dialog History, H and the follow-up question Q, the model needs to predict a set of candidate answers A. The right side of the figure provides visualisation of the task.

The visual dialog task requires an agent to be able to respond to a conversation in the context of a visual input in terms of an image. The idea can be thought of as a visually impaired user having a conversation with an AI agent. The concept of visual dialog task has been posed in two different variants. In the first variant, during training, we train a single agent, that is provided a history of previous rounds of conversation (9 rounds) consisting of question and answers. In the end, the agent is asked a question and has to answer it as a classification task. This consists of training a single bot with an image, a conversation history and a query as input and the task is to classify the output answer as shown in Figure 3. The test setting is similar. In the second variant, there is a question generating bot 'Q-bot' and an answering bot 'A-bot' both of which

are trained. There is a reinforcement learning game between Q-bot and A-bot where they play various rounds of question and answer. At the end, the Q-bot has to guess the image that the A-bot is referring to from a set of images. We have considered the uncertainty estimates for both these models as shown in Figure 2.

Visual dialog task is introduced by [10]. The visual dialog task is defined as, given image I, a caption C, a dialog history till $t-1$ rounds, $H = \{C, (q_1, a_1), ...(q_{t-1}, a_{t-1})\}$ and the following question q_t at round t. The objective of the visual dialog agent is to predict a natural language answer to the question q_t as shown in Figure-3 . The visual dialog problem can be solved into two possible ways; one is by using a generative model and the other by using a discriminative model. In a generative model, given the embeddings of image, history, and question(q_t), a generative model is trained to maximize the likelihood function to predict ground truth answer sequence. The discriminative model receives embedding of an image, history, and question(q_t) along with 100 candidate answers $A_t = \{a_t^1, ...a_t^{100}\}$ and effectively learns to rank the list of candidate answers.

One aspect of previous approaches tends to be a lack of diverse generations of answers; for instance, the tendency to correlate the animal 'zebra' with black and white stripes. In contrast, during conversations, a conversation is interesting if an unexpected or novel observation is raised. In our method, we hope to produce such insights. To do that we need an ability to characterize the space of all possible answers. We do that by using a Gaussian prior for the generation of answers. This allows us to generate samples of plausible answers. We then further use a diversity loss that would penalize correlations between the multiple samples. The final task then lies to choose an appropriate retort or response. To do so, we rely on minimizing uncertainty while generating the answer. We now consider the proposed approach in detail in method section.

2. Related Work

A novel problem, 'Visual dialog' has been introduced recently [9]. In this problem, we aim to answer a question given the context of a conversation about an image. This is one of the latest challenging problems that has been posed in the field of vision and

language. There have been various other related problems considered in the field of vision and language. One of the earliest such problems is that of image captioning where we aim to generate a sentence describing an image [8, 13, 14]. Further, the community moved on to answer questions based on an image in the visual question answering (VQA) task [15, 5]. Recently [16, 17, 18, 19] proposed attention based method to solve this task. Another interesting problem that has been addressed is that of visual question generation (VQG) [20, 21, 22, 23], where the aim is that given an image to generate natural questions similar to that asked by humans. The solution of the visual dialog problem builds up on the previous work conducted for solving the various problems described above.

Visual dialog task requires the agents to have meaningful dialog conversation about the visual content. This task was introduced by Das *et al.* [10]. The authors have proposed three approaches, namely late fusion in which all the history rounds are concatenated, attention-based hierarchical LSTM which handles variable length history and memory-based method for performing results best in terms of accuracy for solving this task. Following up, [24, 25] have proposed generator and discriminator based architecture. Of these, Lu *et al.* [24] consider an attention based method to combine all history rounds to get a single representation. Further works [26, 27, 9] have proposed visual dialog as an image guessing game. The latest work on the visual dialog to obtain state of the art results has been proposed by Jain *et al.* [28]. This work is based on discriminative question generation and answering. In another work, Jain *et al.* [21] have proposed a method to bring diversity in the question generation from an image using Variational Auto-encoder (VAE). Wang *et al.* [29] have proposed a similar kind of method to generate a caption from an image using VAE. In related works, [30] have captured diversity in the caption generation from an image using generative adversary network. Zhang*et al.*[31] proposed a multi-level attention method to fuse different modality in visual dialog task. In contrast to these earlier works, in our framework we consider a fully probabilistic framework for solving the task of visual dialog.

We use Bayesian CNN in our work for obtaining probabilistic image representations. Modeling distribution over CNN filters is still a difficult task. Due to the large number of parameters to be inferred, the posterior distribution becomes intractable. To

approximate this posterior distribution, the variational inference is one of the existing approaches introduced by [32, 33]. Gaussian distribution is the simplest variational approximation used to fit the model parameters to the true distribution of parameters, but it is computationally expensive [33]. This can be overcome using Bernoulli approximation.

There has been some work done in terms of estimating uncertainty in the predictions using deep learning — the work by [34] estimates the predictive variance of the deep network with the help of dropout [35]. [36] has proposed a method to capture model uncertainty for image segmentation task. They observed that softmax probability function approximates relative probability between the class labels, but does not provide information about the model's uncertainty. Recently, [37] has decomposed predictive uncertainty into two major types, namely aleatoric and epistemic uncertainty, which capture uncertainty about the predicted model and uncertainty present in the data itself. Liu *et al.* [38] has proposed probabilistic model for generative network of activity recognition task. This model estimates diversity and uncertainty for the task. The main difference from our method is, the use of a reverse uncertainty attention map (RUAM, detail explanation present in section-4.4). Our method uses gradient certainty attention to improve the attention map, which improves diversity in the generating answer. Here, our objective is to generate diverse answer, to analyze and minimize the uncertainty in answer data, and to analyze the uncertainty of the model for the challenging visual dialog task. We build up on the techniques proposed in several such works to obtain a fully probabilistic framework for solving the visual dialog problem. In the next section we consider the background in terms of Bayesian modeling required for obtaining our probabilistic framework.

It is interesting to see the use of probabilistic models in other domains such as medical applications. Huang *et al.*[39] has proposed a probabilistic topic model for clinical risk stratification. This method recognizes a clinical state of the patient probabilistic fashion with the help of Latent Dirichlet Allocation (LDA). Vulic*et al.* [40] has explained methodology and application of Probabilistic topic modeling in multilingual settings. This method uses bilingual LDA (BiLDA), which is an extension of the LDA model for multilingual settings. Further, Jiang *et al.* [41] has proposed a knowledge

graph based probabilistic method for medical diagnosis. This method uses Boltzmann machines and a Markov network to learn the joint probability distribution.

In our approach we are able to provide two main contributions by using Bayesian deep learning over other existing methods. They are a) we are able to provide uncertainty estimates that allow us to know when a model is not certain about its prediction. It is known that the softmax output of a deep learning system does not reflect the probability of it being correct. By using principled probabilistic models, we are able to provide improved performance along with uncertainty estimates that reflect when the model is not certain about its output and b) We are able to provide more diverse output predictions. Most deep learning techniques that generate outputs are not able to solve and provide diverse generations. Through our approach we are able to solve these aspects by using Bayesian deep learning

3. Background: Bayesian Approach of Model

Consider the distribution $p(x, y)$ over the input features x and labels y. For the visual dialog classification task, x corresponding to joint encoding feature of image, history and query question and y answer class label. For the given observation, X and its corresponding output Y. In a Bayesian framework, the predictive uncertainty of the classification model $(p(y^*|x^*, D))$[1] is trained on a finite set of training data $D = \{x_i, y_i\}_{i=1}^N$. The predictive uncertainty will result in data(aleatoric) uncertainty and the model(epistemic) uncertainty. The model estimates two kinds of uncertainty, i.e., data uncertainty and model uncertainty. The posterior distribution describes the data uncertainty over class labels, given set of model parameters w, and the model uncertainty is described by the posterior distribution over model parameters w, given input data. The predictive uncertainty for new example point x^* by integrating over all possible set of parameters w is given by

$$p(y^*|x^*, X, Y) = \int \underbrace{p(y^*|x^*, w)}_{Data} \underbrace{p(w|X, Y)}_{Model} dw \tag{1}$$

[1] Standard shorthand notation for $p(y = y^*|x^*, X, Y) = p(y^*|x^*, D)$

Our main objective is to find the best set of weights of our model that will generate our data X, Y. One of the approaches to make Bayesian inference is to compute the posterior distribution overweights, i.e., $p(w|X,Y)$. This distribution captures the best plausible set of model parameters given our observed data.

$$p(w|X,Y) = p(Y|X,W)p(W)/p(Y|X)$$

It is challenging to perform inference over the Bayesian network because the marginal probability $p(Y|X)$ of the posterior cannot be evaluated analytically. So, the posterior distribution $p(w|X,Y)$ is intractable. To approximate the intractable posterior distribution, various approximation approaches are proposed in [42, 34, 33]. Variational inference is one of the approximating technique, where the posterior $p(w|X,Y)$ is approximated by a simple distribution $q_\theta(W)$, where θ is the parameterized by variational parameter $p(w|X,Y) \approx q_\theta(W)$. We thus minimize the Kullback–Leibler(KL) divergence between approximate distribution $q_\theta(w)$ and the posterior $p(w|X,Y)$ w.r.t θ, which is denoted by $KL(q_\theta(w)||p(w|X,Y))$.

$$\begin{aligned} KL(q_\theta(w)||p(w|X,Y)) &\propto -\int q_\theta(w)\log p(Y|X,w)dw + KL(q_\theta(w)||p(w)) \\ &= -\sum_{i=1}^{N}\int q_\theta(w)\log p(y_i|f^{\hat{W}}(x_i))dw + KL(q_\theta(w)||p(w)) \end{aligned}$$

Minimizing the KL divergence is equivalent to maximizing the log evidence lower bound [42] with respect to the variational parameters defining $q_\theta(w)$,

$$L_{VI} = \int q_\theta(w)\log p(Y|X,w)dw - KL(q_\theta(w)||p(w)) \tag{2}$$

The intractable posterior problem i.e., averaging over all the weight of the BNN, is replaced by the simple distribution function. Now we need to optimize the parameter of simple distribution function instead of optimizing the original neural network's parameters. Furthermore the integral in equation 1 (predictive posterior) is also intractable for the neural network, which is approximated via sampling using Monte Carlo dropout [34] or Langevin Dynamics [43] or explicit ensembling [44]. So we approximate the

integral with Monte Carlo sampling.

$$
\begin{aligned}
p(y^* = c|x^*, X, Y) &= \int p(y^* = c|x^*, w)p(w|X, Y)dw \\
&\approx \int p(y^* = c|x^*, w)q_\theta(w)dw \\
&\approx \frac{1}{M}\sum_{i=1}^{M} p(y^* = c|x^*, w^{(i)})q(w^{(i)})
\end{aligned}
\tag{3}
$$

where $w^{(i)} \sim q(w^{(i)})$, which is modeled by the dropout distribution and M samples of $w^{(i)}$ is obtained. , each $p(y^*|x^*, w^{(i)})$ in an ensemble $p(y^*|x^*, w^{(i)})_{i=1}^{M}$ obtained sampled from $q(w^{(i)})$. In the following section, we have discussed Bayesian CNN and Bayesian LSTM.

3.1. Bayesian CNN

One way to define a Bayesian neural network [34] is to place a prior distribution over neural network weights, $w = (W_i)_{i=1}^{L}$. Given weight matrix W_i and bias b_i for ith layer, we use standard Gaussian prior distribution over the weight matrix $p_0(W_i) = \mathcal{N}(W_i; 0, 1)$. The variational Bayesian approximation in a Bayesian neural network can be interpreted as adding stochastic regularization in the deterministic neural network. The stochastic regularization technique is equivalent to multiplying random noise ϵ_i with neural network weight matrices M_i.

$$
\begin{aligned}
&W_i = M_i \cdot diag([\epsilon_{i,j}]_{j=1}^{K_i}) \\
&\epsilon_{i,j} \sim Bernoulli(p_i), i = \{1, ., L\}, j = \{1, ., K_{i-1}\}
\end{aligned}
\tag{4}
$$

where, $\epsilon_{i,j}$ is a Bernoulli distributed random variable with probability p_i. The $diag(.)$ operator maps vectors to diagonal matrices, whose diagonal elements are the elements of the vectors. The set of variational parameters M_i is now the set of matrices $\theta = \{M_i\}_{i=1}^{L}$. The binary variable $\epsilon_{i,j} = 0$ indicates the corresponding element j in the layer $i-1$ is dropped out as an input to layer i. In CNN with dropout [45], the forward

propagation is formulated as,

$$\begin{aligned} m_k^i &\sim Bernoulli(p_i) \\ \hat{a}_k^i &= a_k^i * m_k^i \\ z_j^{i+1} &= \sum_{k=1}^{n^{(l)}} Conv(W_j^{l+1}, \hat{a}_k^i) \end{aligned} \tag{5}$$

Here a_k^i denotes the activations of feature map k ($k = 1, 2, \ldots, n^{(l)}$) at layer l. The mask matrix m_k^l consists of independent Bernoulli variables $m_k^l(i)$. This mask is sampled and multiplied with activations in kth feature map at layer l, to produce dropout-modified activations $\hat{a}_k^i$. These modified activations are convolved with filter W_j^{l+1} to produce convolved features z_j^{i+1}. The function f is applied element wise to the convolved features to get the activations of convolutional layers.

3.2. Bayesian LSTM

The conventional LSTM is a neural network that maps LSTM state s_t (at time step t) and input x_t to a new LSTM state s_{t+1}, $f_l : (s_t, x_t) \rightarrow s_{t+1}$. The state of LSTM is given by $s_t = (c_t, h_t)$, where c_t is a memory state and h_t is the output of the hidden state. To train a input sequence of length T, $x_1, x_2..., x_T$, the LSTM cell is unrolled T times in to a feed forward network with initial state s_0 and can be represented by

$$s_j = f_l(s_{j-1}, x_j)$$

In Bayesian LSTM, let $p(y^*|w, x^*)$ be the likelihood of the neural network, then the posterior of the network is approximated to $q(w)$ by minimizing the variational free energy $L(w)$[46, 47].

Minimizing the variational free energy is equivalent to maximizing the likelihood $\log p(y|x, w)$ subject to KL divergence, which regularizes the parameters of the network.

$$L(w) = -E_{q(w)}[\log p(y^*_{1:T}|x^*_{1:T}, w)] + KL(q(w)||p(w))$$

Here $\log p(y_{1:T}|x_{1:T}, w)$ is the likelihood function of the sequence and the expectation in the previous equation is approximated by the Monte Carlo sampling. The

predictive posterior for LSTM is calculated just as in equation 3 by,

$$
\begin{aligned}
p(y^*_{1:T}|x^*_{1:T}, X, Y) &= \int p(y^*_{1:T}|x^*_{1:T}, w)p(w|X, Y)dw \\
&\approx \frac{1}{M}\sum_{m=1}^{M} p(y^*_{1:T}|x^*_{1:T}, \hat{w}_m)dw
\end{aligned} \tag{6}
$$

with $\hat{w}_m \sim q_\theta(w)$, where $q_\theta(w)$ is called the dropout distribution for LSTM.

In general, there are two broad categories of uncertainty [34] i.e. Epistemic Uncertainty and Aleatoric uncertainty. Epistemic Uncertainty captures the uncertainty present in the model. This means, the uncertainty in our model where it cannot explain some data as it has not been trained well due to lack of data. This can be reduced by observing more data. The uncertainty present in data is captured by aleatoric uncertainty. This captures inherent uncertainty in data due to cases such as occlusion, sensor noise etc. This thus captures uncertainty that our data cannot explain. One part of aleatoric uncertainty known as *heteroscedastic* aleatoric uncertainty tries to capture the fact that parts of the observation space may have more noise. For instance, if we know that answering certain kinds of questions is more ambiguous than answering other kinds of questions. Better estimation and minimizing this uncertainty specifically aids us in improving performance in our task.

4. Methods

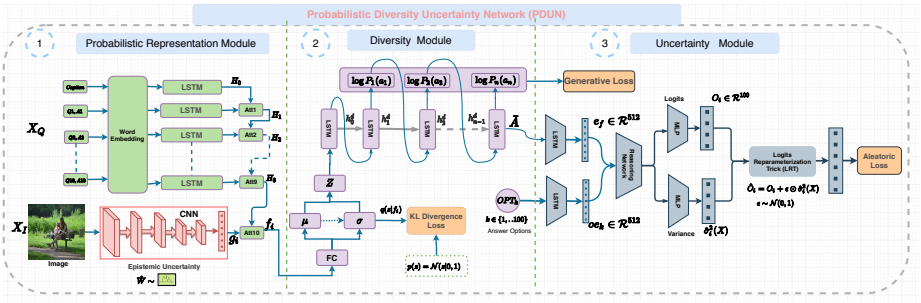

Figure 4: Probabilistic Diversity Uncertainty Network(PDUN), Bayesian CNN/LSTM is used to obtain the embeddings g_i, f_i, h_i which is then fused using the Fusion Module to get e_f. Then correlation is found between fused embedding with answer option embedding. Finally, variance and logits output are obtained using MLP, which is then used in Logits Reparameterization Trick(LRT) to get final softmax output.

4.1. Overview

Our method consists of three parts viz. 1) Probabilistic Representation Module, 2) Latent feature-based Diverse Answer Generation Module, and 3) Uncertainty Module as illustrated in figure 4 :

1. **Probabilistic Representation Module:** We obtained probabilistic embeddings for various inputs such as image, question and history using Bayesian CNN and RNN modules. We term this as a 'representation module'. Then, we attended at each round of question with previous history and obtained new history for next round. Finally we attended image embedding with history context to obtain Context embedding.
2. **Latent feature-based Diverse Answer Generation Module:** We obtained a latent vector representation from mean and variance of the context vector using a standard re-parameterization trick [48] that is commonly used for most variational methods. After that, we generate a diverse set of answers from the latent vector.
3. **Uncertainty Module:** We measure both epistemic and aleatoric uncertainty and we then use reverse uncertainty technique to minimize the uncertainty present in the attention maps.
4. Now, the obtained answers make a reasoning with each answer option and two outputs are obtained, one 100 dimensional output prediction vector and the other one is an uncertainty vector.

4.2. Probabilistic Representation Module

We adopt a probabilistic representation module that has been previously considered in Patro *et al.* [11] for the VQA task. However, using this representation for visual dialog requires us to also consider the history of previous dialogs as a new input. By using a probabilistic representation, we are able to investigate the uncertainty in any part of the full proposed model. To obtain this, we use the methods of Bayesian CNN and Bayesian LSTM that has been discussed in the background section. Given an input image x_i, we obtain an image embedding g_i by using a Bayesian CNN that we parameterized through a function $G(x_i, W_i)$, where W_i are the weights of the Bayesian CNN.

We extract $g_i \in \mathcal{R}^{w \times h \times c}$ dimensional CONV-5 feature from Bayesian CNN network as shown in figure 4. We obtain g_q, g_h encoding feature for given question and history, after passing through an LSTM (Bayesian LSTM Network), which is parameterized using the function $G_q(X_{WE}, \theta_l)$, where θ_l are the weights of the LSTM as shown in figure 4. Similarly, we obtain answer embedding G_a parameterised by $G(X_a, \theta_a)$. After this, the question and answer embedding are combined to obtain a history embedding. To model the Bayesian CNN [34], we use pre-trained CNN layers and put dropout layer with dropout rate p, before each CNN layer. Similarly for Bayesian: LSTM [47], we add dropout on each input and a hidden layer of the LSTM cell. These are input to an attention network that combines question-answer pair with previous history embedding using a weighted softmax function and produces a weighted output attention vector g_f. There are various ways of modeling the attention network. In this paper, we have evaluated the network proposed in SAN [49]. In the last round, we combine image embedding with the last history embedding to get a dialog context vector. At each round, we attend over the question representation with the previous history (combined question-answer representation). In the first round, the previous history is an encoded caption feature. In the final round, we attend to image representation with the appropriate history representation to obtain an attentive encoder feature, g_f. The attention mechanism is illustrated as follows:

$$\begin{aligned} g_a &= \tanh(W_c g_i + W_q(g_q || g_h) + b_c) \\ \alpha &= \text{Softmax}(W_a g_a + b_a) \end{aligned} \tag{7}$$

where $||$ means concatenation, W_a, W_c, W_q, b_c, b_a are the weights and bias of different layers.

4.3. Latent feature based Diverse Answer Generation Module

This module mainly focuses on representing a latent representative vector from attentive encoder module and generate a diverse answer using answer generator. We use the VAE [48] based generative framework to generate diverse answer from the attentive encoder. We obtain mean, $\mu = W_\mu g_f$ and log variance, $\log \sigma^2 = W_\sigma g_f$, where μ and σ are the parameters of a multivariate Gaussian distribution. We train

this network to learn a variational distribution which is close to a prior defined by the normal distribution with zero mean and unit variance i.e., $\mathcal{N}(0, 1)$. Then we obtain a latent vector representation $\mathbf{z}$ by using the reparameterization trick $\mathbf{z} = \mu + \epsilon \odot \sigma$. The major concern for answer generation is the spread of the variance in the latent representation. Our main objective is to increase the spread in the Gaussian as much as possible for generating diverse answers. If the spread of the Gaussian is too low (~ 0), then we have sampled similar answers, and if the variance is too high ($\sim \infty$), then we have sampled from a uniform distribution. Hence, we want to put some constraints on variance such that our sampled latent representations are as diverse as possible. A diversity loss which minimizes the correlation between the latent representations is introduced to ensure this. Let us define z_1 and z_2 as the two latent vectors randomly sampled from the $\mathcal{N}(\mu, \sigma)$, $z_1 = \mu + \epsilon_1 \odot \sigma, z_2 = \mu + \epsilon_2 \odot \sigma$, where ϵ_1 and ϵ_2 are sampled from $\mathcal{N}(0, 1)$ and μ and σ are Gaussian parameters. The diversity loss is given by

$$L_{diverse} = \frac{\langle (z_1 - \alpha), (z_2 - \alpha) \rangle}{max(||z_1 - \alpha||_2 \cdot ||z_1 - \alpha||_2, \gamma)} \tag{8}$$

Where $\gamma = 10^{-8}$ is used to avoid division by zero, and α is the average of all the z samples. Similarly, we obtain an average loss for k sample points (in our experiments, we choose k to be 100) randomly sampled from the latent distribution. This loss ensures that these latent vectors are as far as possible.

Finally, the diverse latent feature is input to an LSTM based answer decoder module to generate diverse answers. The softmax probability for the predicted answer token at different time steps is given by the following equations:

$$h_0 = Z_i = \mathcal{N}(\mu, \sigma)$$
$$x_t = W_e * a_t, \forall t \in \{0, 1, 2, ...T-1\}$$
$$h_{t+1} = \text{LSTM}(x_t, h_t), \forall t \in \{0, 1, 2, ...T-1\}$$
$$\hat{y}_{t+1} = softmax(W_o * h_{t+1})$$
$$L_{CE} = -\frac{1}{C} \sum_{j=1}^{C} y_j \log \mathrm{p}(\hat{y}_j | f_o)$$

where $\hat{y}_{t+1}$ is the predicted answer class and f_o is the context. Now, we classify the generated diverse answer among 100 classes in order to rank them with 100 ground

truth answer options. We use a reasoning network to perform reasoning by predicting an answer and comparing it with the ground truth answer to obtain the final score. A similar approach has been used by Das *et al.* [9].

4.4. Uncertainty Representation Module

Through the previous module, we obtain the ability to generate diverse answers. The task then is to choose an appropriate answer that is correct. To do that, we use an uncertainty representation module that characterizes the uncertainty among the diverse set of candidate answers (i.e., the classes present in the answers). We want to be certain about the response that is chosen. That is, we would like to minimize the uncertainty. We do that by using an explicit loss for reducing uncertainty.

In this work, we also incorporate the attention regions, which specifies the spatially distributed weights given to a specific region embedding while generating the answer. To obtain the best embedding, we consider the ground-truth answer and through attention, consider the corresponding spatial location. This region is multiplied with the uncertainty for generating the spatial attention weighted uncertainty corresponding to the ground-truth answer. We want to increase the weight for the spatial attention corresponding to generating the ground-truth answer and minimize the uncertainty for the same. At the same time, we would like to increase the uncertainty for all other answers and minimize the weight given in terms of attention to all other regions. We achieve this through a reverse uncertainty based attention map (RUAM) that is shown in figure 5.

Reverse Uncertainty based Attention Map (RUAM): Patro *et al.*[11] have proposed a model to estimate aleatoric and predictive uncertainty for Visual Question Answering task, where the gradient of uncertainty loss and gradient of classification is multiplied to improve attention feature. Kurmi *et al.* [50] have also proposed a similar kind of network in the domain adaption task, where they train the discriminator network to reduce uncertainty in source and target domain. We follow a similar type of network in a visual dialog task to reduce uncertainty in the attention mask with the help of a predicted answer in the dialog turns. We stress more on those attention regions whose uncertainty is less and vice-versa. The aleatoric uncertainty occurs due to

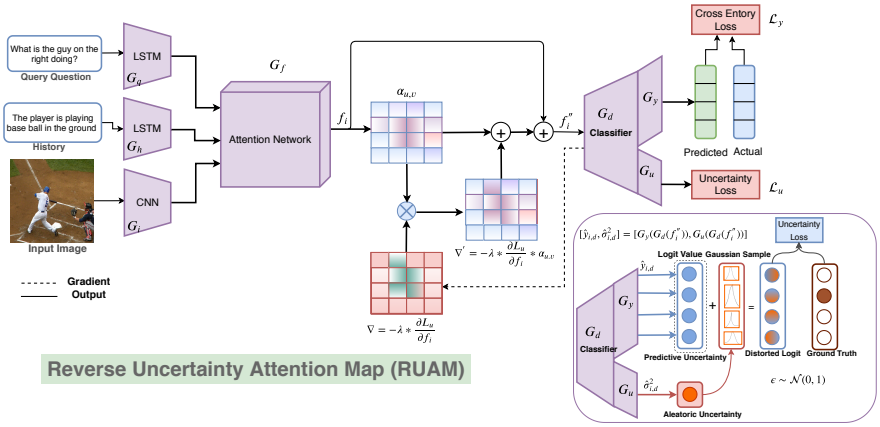

Figure 5: Reverse Uncertainty based Attention Map (RUAM): We obtain attention embedding f_i from the attention network G_f using image, question and history embeddings g_i, g_q, g_h. Then we classify into answer class and obtain the uncertainty present in the data. Then we obtain reverse uncertainty map with will combine with attention map to get better confidence on the attention map as shown in the figure.

corruption in the feature or noise in the attention regions. These regions are the main source of predicting the wrong answer in the visual dialog.

We adopt a Bayesian framework to predict answer classification uncertainty efficiently. We make our answer classifier as Bayesian and perform probabilistic inference over the classifier to obtain the final answer score. We adopt a Bayesian classifier as considered in several works [46, 34, 47, 50, 11]. The Bayesian classifier is obtained by applying dropout after every fully connected (FC) layer in the classifier and the Bayesian classifier predicts answer class logits $y_{i,d}$ and aleatoric uncertainties. These are obtained as follows:

$$y_{i,d} = G_y(G_d(f_i)), \quad \sigma_{i,d} = G_v(G_d(f_i)) \tag{9}$$

where f_i is the attention feature for input image sample x_i, question sample x_q and history sample x_h, which is obtained by the attention feature extractor $G_f : f_i = G_f(G_i(x_i), G_q(x_q), G_h(x_h)$, where G_y and G_v are the logits and aleatoric variance predictor of the classifier G_d respectively. The whole model is trained with the help of uncertainty loss (More details are present in 4.5) and cross-entropy loss. The uncertainty loss helps the classifier to make the classifier features more robust for prediction.

Finally, we measure the uncertainty for our answer prediction and found it to be lower.

We learn and estimate observational noise parameter $\sigma_{i,d}$ to capture the uncertainty present in the input data (Image, History and Question). This can be achieved by corrupting the logit value ($y_{i,d}$) with the Gaussian noise with variance $\sigma_{i,d}$ (diagonal matrix with one element for each logits value) before the softmax layer. We used a Logit Reparameterization Trick (LRT) [51], which combines two outputs $y_{i,d}, \sigma_{i,d}$ and then we obtain a loss with respect to ground truth. That is, after combining we get $\mathcal{N}(y_{i,d}, (\sigma_{i,d})^2)$ which is expressed as:

$$\hat{y}_{i,t,d} = y_{i,d} + \epsilon_{t,d} \odot \sigma_{i,d}, \quad where \quad \epsilon_{t,d} \sim \mathcal{N}(0,1) \tag{10}$$

$$\mathcal{L}_u = \sum_i \log \frac{1}{T} \sum_t \exp\left(\hat{y}_{i,t,M} - \log \sum_{M^{'}} \exp \hat{y}_{i,t,M^{'}}\right) \tag{11}$$

where $M^{'}$ is a discrete word token present in each sample sentence. $y_{i,t}$. M is a discrete word token present in the ground truth sentence, $\mathcal{L}_u$ is the uncertainty loss function, and t $\in$ T is the number of Monte Carlo simulations. $\sigma_{i,d}$ is the standard deviation, ($\sigma_{i,d} = \sqrt{v_{i,d}}$).

Now, we obtain uncertainty for attention map $\alpha_{att} \in \mathcal{R}^{u \times v}$ of width u and height v using following steps such as, we first compute gradient of the predictive uncertainty σ_g^2 of our generator with respect to the features f_i. This gradient of the uncertainty loss $\mathcal{L}^u$ with respect to the attention feature f_i is given by $\frac{\partial \mathcal{L}_v}{\partial f_i}$. Now we pass the uncertainty gradient through a gradient reversal layer to reverse the gradient to get certainty mask for the attention map. This is given by

$$\nabla_u = -\gamma * \frac{\partial \mathcal{L}_u}{\partial f_i}$$

We perform an element-wise multiplication of the forward attention feature map and reverse uncertainty gradients to get an enhanced attention feature map i.e.

$$\alpha^{'}_{u,v} = -\gamma * \frac{\partial \mathcal{L}_u}{\partial f_i} * \alpha_{u,v} \tag{12}$$

The positive sign of the gradient γ indicates that the aleatoric certainty is activated on these regions and vice-versa. We apply a ReLU activation function on the product of

gradients of the attention map and the gradients of aleatoric certainty as we are only interested in attention regions that have a positive influence for a corresponding answer class, i.e. attention pixels whose intensity should be increased in order to increase y^c, where negative values are multiplied by γ (large negative number). Negative attention pixels are likely to belong to other categories in the image.

$$\alpha^{''}_{u,v} = ReLU(\alpha^{'}_{u,v}) + \gamma ReLU(-\alpha^{'}_{u,v}) \tag{13}$$

Images with higher aleatoric uncertainty correspond to lower certainty. Therefore the certain regions of these images should have lower attention values. We use residual connection to obtain the final attention feature by combining original attention feature with the reverse uncertainty map $\alpha^{''}_{u,v}$. This is given by:

$$\begin{aligned} \alpha_{new} &= \alpha_{u,v} + \alpha^{''}_{u,v} * \alpha_{u,v} \\ f^{'}_{i} &= \sum_{u,v} g_i * \alpha_{new} \end{aligned} \tag{14}$$

Where, $g_i \in G_i(x_i)$. The final attention feature $(f^{''}_i)$ can be obtained by combining attention feature (f_i) with RUAM based new attention feature $(f^{'}_i)$.

$$f^{''}_i = f_i + f^{'}_i \tag{15}$$

We show here, that using reverse uncertainty based attention Map (RUAM) results in an improved attention network and the attention confidence also increases. The entropy and predicted variance of the sampled logit's probability can be calculated as:

$$H(\hat{y}_{i,t}) = -\sum_{m=1}^{M} p(\hat{y}_{i,t} = M) * \log p(\hat{y}_{i,t} = M) \tag{16}$$

The predictive uncertainty is the combination of entropy and variance of T sample outputs (of randomly masked model weights).

$$\sigma_p^2 = \frac{1}{T}\sum_{t=1}^{T} H(\hat{y}_{i,t}) + \frac{1}{T}\sum_{t=1}^{T} v_{i,t}^2 \tag{17}$$

Where $H(\hat{y}_{i,t})$ is the entropy of the probability $p(\hat{y}^c_{i,t})$, which depends on the spread of the probabilities and the variance captures both the spread and the magnitude of

outcome values $\hat{y}_{i,t}$. Algorithm-1 explains details about reverse uncertainty map for attention mask.

LSTM cell in the uncertainty module:We use same Bayesian RNN module as a decoder LSTM as introduced in the background section. To measure uncertainty over LSTM cell, we predict a 100 dimensional output vector. We then took the gradient of LSTM cell with respect to the output vector and then obtain gradient of attention cell with respect to LSTM cell [47]. To measure uncertainty we use the same procedure. We expand the output of all the time steps to single vector. We then measure aleatoric variance using variance network which converts logit vector dimension to single unit to measure uncertainty. Then we add Gaussian noise of logit variance and added to original logit value and it is called as a distorted output. We then take categorical cross entropy loss between distorted output with ground truth output. This loss is then minimized.

4.5. Cost Function

Finally, we trained our complete PDUN model with the help of answer generation loss and uncertainty loss. The answer generation loss L_{gen} is the combination of cross entropy loss L_{CE}, to generate each and every token in the answer sequence, KL divergence loss L_{KL}, to bring the approximate posterior closer to $\mathcal{N}(0, 1)$, and the diversity loss L_{div} 8, to ensure diverse answer generation. The cost function used for obtaining the parameters θ_f of the attention network, θ_c of the classification network, θ_y of the prediction network and θ_u for uncertainty network is as follows:

$$C(\theta_f, \theta_c, \theta_y, \theta_u) = \frac{1}{n}\sum_{j=1}^{n} L_y^j(\theta_f, \theta_c, \theta_y) + L_{KL}^j(\theta_f, \theta_c) + L_{div}^j(\theta_f, \theta_c) + \eta L_u^j(\theta_f, \theta_c, \theta_u)$$

where n is the number of examples, and η is a hyper-parameter that is fine-tuned using validation set and L_c is standard cross entropy loss. We train the model with this cost function till it converges so that the parameters $(\hat{\theta}_f, \hat{\theta}_c, \hat{\theta}_y, \hat{\theta}_u)$ deliver a saddle point function

$$(\hat{\theta}_f, \hat{\theta}_c, \hat{\theta}_y, \hat{\theta}_u) = \arg \max_{\theta_f, \theta_c, \theta_y, \theta_u} (C(\theta_f, \theta_c, \theta_y, \theta_u)) \tag{18}$$

Algorithm 1 Reverse Uncertainty based Attention Map (RUAM)

procedure RUAM(I, Q, H)

Input: Image X_I, Question X_Q, History X_H

Output: Answer y

while loop **do**

Attention features $G_f(G_i(X_I), G_q(X_Q), G_H(X_H)) \leftarrow f_i$

Answer Logit $G_y(G_d(f_i)) \leftarrow \hat{y}$

Data Uncertainty $G_u(G_d(f_i)) \leftarrow \sigma_A^2$

$\sigma_W^2 = \sigma_A^2 + H(\hat{y}_{i,t})$, (Ref: eq- 4)

Ans cross entropy $\mathcal{L}_y \leftarrow$ *loss*$(\hat{y}, y)$

Variance Equalizer [52] $\mathcal{L}_{VE} := \sum ReLU(\exp^{\sigma_w^2} - \exp^{I})$,

while $t = 1 : \#MC - Samples$ **do**

Sample $\epsilon_t^w \sim \mathcal{N}(0, \sigma_W^2)$

Distorted Logits: $\hat{y}_{i,t} = \epsilon_t^w + \hat{y}_i$

Gaussian Cross Entropy [52] $L_{GCE} = -\sum y \log p(\hat{y_d}|F(.))$

Distorted Loss : $\mathcal{L}_{UDL} = \exp(\mathcal{L}_y - \mathcal{L}_{GCE})^2$

Aleatoric uncertainty loss $\mathcal{L}_u = \mathcal{L}_{GCE} + \mathcal{L}_{VE} + \mathcal{L}_{UDL}$

Compute Reverse Gradients w.r.t f_i, $\nabla_u = -\lambda * \frac{\partial \mathcal{L}_U}{\partial f_i}$

Certainty Activation for attention $\alpha_{u,v}^{'} = \nabla_u * \alpha_{u,v}$

Certainty Activation for attention $\alpha_{u,v}^{''} = ReLU(\alpha_{u,v}^{'}) + \gamma * ReLU(-\alpha_{u,v}^{'})$

New Attention gradient $\alpha_{new} = \alpha_{u,v} + \alpha_{u,v}^{''} * \alpha_{u,v}$

New attended feature: $f_i^{'} = \sum_{u,v} f_i * \alpha_{new}$

Final attended feature: $f_i^{''} = f_i + f_i^{'}$

update $\theta_f \leftarrow \theta_f - \eta \nabla_y^{'}$

5. Experiments

We evaluate the proposed method in the following steps: First, we evaluate our proposed uncertainty model against other variants described in section 5.2. Second, we have shown analysis results for epistemic uncertainty in figure-6 and aleatoric uncertainty in figure-7 and in table 2. Third, we further analyze effect of noise in aleatoric

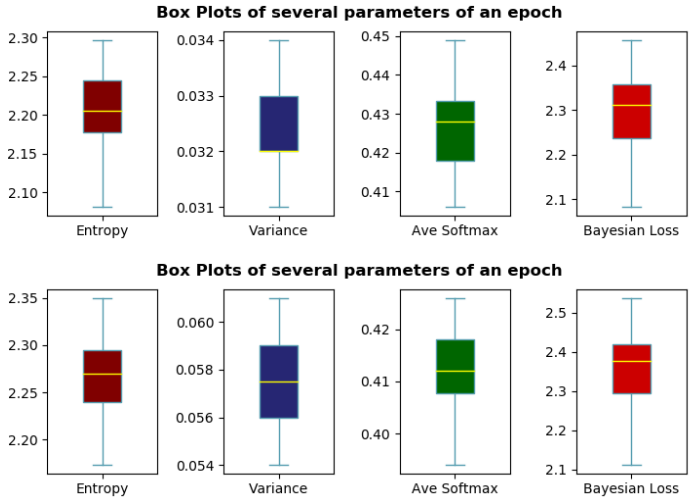

Figure 6: This shows the measurement of Entropy, Variation, Softmax scores and Bayesian loss for Bayesian model with dropout value 0.5 and 0.1 in first and second plot respectively for capturing Epistemic Uncertainty

and epistemic uncertainty in table 3. Fourth, we compare diversity score for different variants of our model in table 5. Fifth, we compare our network with state-of-the-art methods such as 'visdial' [10] in table 4. Then, we have shown the Grad-CAM [53] visualization of activation due to aleatoric uncertainty and baseline model (late fusion). We further compare our network with state-of-the-art methods such as visdial [10] model. Finally, we have provided some qualitative results of our visual dialog model. The quantitative evaluation is conducted using standard retrieval metrics, namely (1) mean rank, (2) recall @k, (3) mean reciprocal rank (MRR) of the human response in the returned sorted list.

5.1. Dataset

We evaluate our proposed approach by conducting experiments on Visual Dialog dataset [10], which contains human annotated questions based on images of MS-COCO dataset. This dataset was developed by pairing two subjects on Amazon Mechanical Turk to chat about an image. One person was assigned the job of a 'questioner' and the other person act as an 'answerer'. The questioner sees only the text description of an image which is present in caption from MS-COCO dataset. The image remains hidden to the questioner. Their task is to ask questions about this hidden image to "imagine

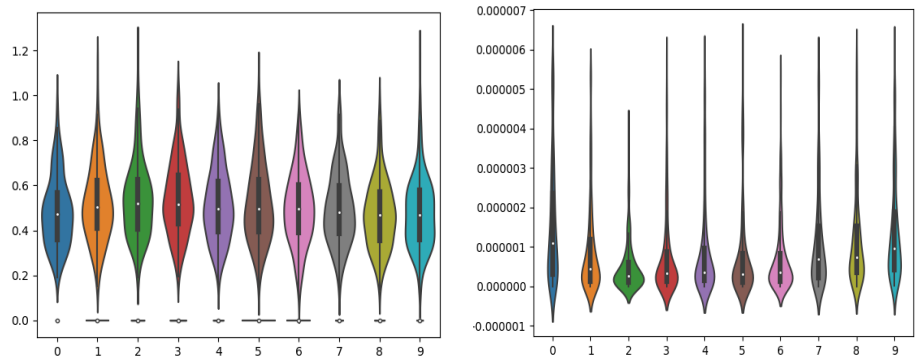

Figure 7: Left Graph: This shows how aleatoric uncertainty loss varies over different turns in visual dialog. There is an eventual decreasing trend. Right Graph: This shows how variance equalizer loss varies over different turns in visual dialog. There is an eventual increasing trend.

Loss	**R1**	**R5**	**R10**	**MRR**	**Mean**
Baseline	40.9	72.4	82.8	0.550	05.95
VE	30.9	51.4	62.5	0.371	12.41
CE	41.1	72.6	82.9	0.556	05.98
GCE	43.8	77.2	88.0	0.587	04.79
CE+VE	42.3	74.6	84.7	0.551	05.50
VE+GCE	44.3	79.5	89.6	0.599	04.41
CE+GCE	45.4	80.6	91.2	0.610	03.95
ACE	47.0	82.4	92.3	0.622	03.81

Table 1: In this table, we show results of all the variant of aleatoric loss on visual dialog-v1.0[10] in test-std dataset. From the table we observe that "GCE" performance best over individual loss, then we see combination with "GCE" is improves in performance and finally when we combine all three individual loss "ACE" performs better compare to other model.

the scene better". The answerer sees the image and caption and answers the questions asked by the questioner. The two of them can continue the conversation by asking and answering questions for 10 rounds at max. We have performed experiments on "VisDial 1.0" version of the dataset. "VisDial v1.0" contains 123k dialogs on COCO-train and 2k on "VisualDialog_val2018" images for val and 8k on "VisualDialog_test2018" for test-standard set. The caption is considered to be the first round in the dialog history.

Type of Uncertainty	Mean	Std
Aleatoric (with CE)	0.0051	08.6772
Aleatoric (with VE)	0.0044	07.3534
Aleatoric (with GCE)	0.0039	03.4317
Aleatoric (with ACE)	0.0032	02.1193
Epistemic (50% training)	0.6680	66.9321
Epistemic (75% training)	0.6310	42.8923
Epistemic (100% training)	0.5520	36.8110

Table 2: In this table we show the changes in uncertainty(Aleatoric and Epistemic uncertainty) measurement score with the help data argumentation. We calculate mean and standard deviation as measurement for both uncertainties.

5.2. Ablation Analysis on Model Parameter for Uncertainty

Aleatoric Cross Entropy consists of distorted (Gaussian Cross Entropy (GCE)), undistorted (Cross Entropy (CE)) loss, and Variance Equalizer (VE) loss.. The first block of the table 1 analyses individual loss function and its comparison is provided in that table. We use these models as our baseline and compare other variations of our model with the best single loss function. The GCE loss performs best among all the 3 losses. This is reasonable as GCE can guide the loss function to minimize the variance in the data. The second block of table 1 depicts the models which uses combination of the loss function as variations of our method such as GCE, VE or CE. We see an improvement of around 4% in R@1, 8% in R@5 score, 9% in R@10 score and 5% in MRR score from the baseline score. The combination of GCE loss and CE performs best among all the 3 cases. The third block takes into consideration all the loss functions ACE (GCE+CE+VE) and we see an improvement of around 6% in R@1, 10% in R@5 score, 10% in R@10 score and 7% in MRR score from the baseline score. The behaviour of dialog turn for a particular example is shown in 7. The first part of the figure 7 shows, how aleatoric uncertainty loss varies over different turns in visual dialog. As dialog progress the width of the dialog turn decreases. There is an eventual decreasing trend. The second part of the figure shows how variance equalizer loss varies over different turns in visual dialog. There is an eventual increasing trend.

We can observe the third and fourth turn is more uncertain. This indicates that to have a successful dialog, it basically depend on the central part of the dialog not only start and end tuns of the dialog.

5.2.1. Analysis of Epistemic Uncertainty

One of the main purposes of the Bayesian deep learning is that it improves both the predictions and the uncertainty estimates of the model. We have measured uncertainty score in terms of mean and variance for all the dialog prediction in "val-v1.0" dataset. We have also measured uncertainty for a single dialog in the dataset. Here, we split our training data into three parts. In first part the model is trained with 50% of the training data. Then, second part is trained by 75% of training data and third part is trained by full dataset as shown in second block of the table 2. It is observed the epistemic uncertainty decreases as training data increases.

Type of Uncertainty	**Mean**	**Std**
Aleatoric (original)	0.0067	08.956
Aleatoric ($\gamma = 0.8$)	0.0123	11.717
Aleatoric ($\gamma = 1.2$)	0.0034	06.353
Epistemic (original)	0.6714	70.213
Epistemic ($\gamma = 0.5$)	0.7017	71.893
Epistemic ($\gamma = 0.8$)	0.6461	69.117

Table 3: In this table, we show change in uncertainty (both aleatoric & epistemic uncertainty) measurement score with noise parameters. We calculate mean & standard deviation as measurement for both uncertainties.

5.2.2. Analysis of Aleatoric Uncertainty

Here, we have captured data uncertainty by checking contribution of each terms in aleatoric uncertainty as shown in first block of the table 2. From the measurements, it can be easily seen that comparing aleatoric uncertainty of an image with epistemic uncertainty of another image doesn't make sense because of significant difference in their values. But both the uncertainties can be separately compared for different images to see which answer is more uncertain.

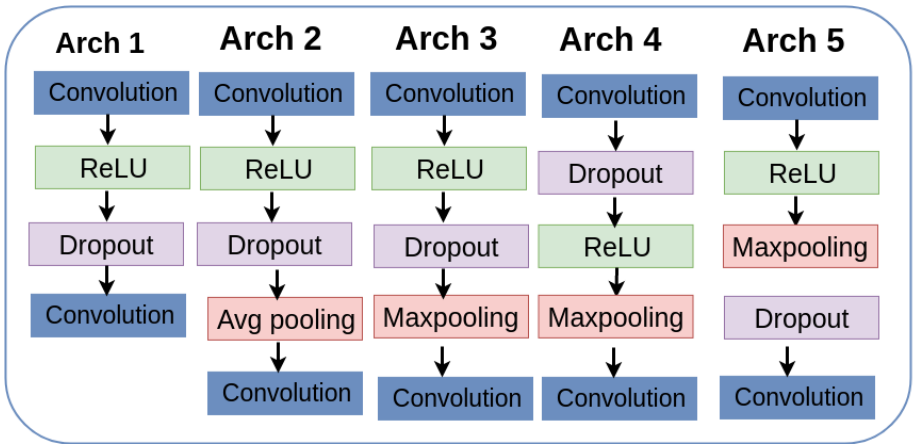

Figure 8: Bayesian CNN experiment based on dropout, Max-pooling and average pooling. We found out that Architecture 5 works best and we used it throughout our experiments.

5.2.3. Analysis of Noise in Aleatoric & Epistemic uncertainty

We have performed another ablation study for change in uncertainty based on noise value. We estimated both aleatoric and epistemic uncertainty for visual dialog dataset. We randomly selected 200 examples on val dataset and applied noise to the image and question responses and observed that uncertainty value changes on seeing noise image and noise question. So we applied noise value γ of 0.8 to decrease pixel value and $\gamma = 1.2$ to increase pixel value i.e. there is inverse proportionality. Mean and standard deviation of uncertainties are recorded in the table. From table 3, it can be observed that aleatoric uncertainty is very small as compared to epistemic uncertainty. The aleatoric uncertainty changes much rapidly as noise changes in comparison to that of epistemic uncertainty.

5.2.4. Analysis of Epistemic Uncertainty : Dropout

We experimented with various dropout ratios and use the following values for the same. For implementing Bayesian CNN, we used dropout ratio of (0.1, 0.2, 0.3, 0.4, 0.5) for each stack of convolutional layers respectively and 0.5 for FC layers. As the number of neurons increase in subsequent layers, we increase the dropout ratio for better generalization. For Bayesian LSTM, we use dropout ratio 0.3 for input & hidden layers and for output layer we have used 0.5 dropout ratio similar to [46]. We further

(a)Distorted loss(GCE) (b) Undistorted loss(CE) (c) Variance Equalizer loss(VE)

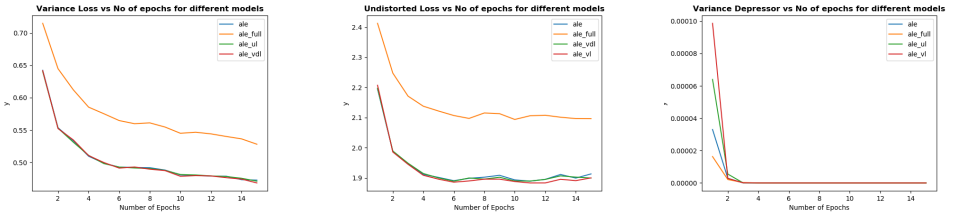

Figure 9: This figure shows role of different types of Losses over Epochs. From the plot we observed that variance is decreasing as it goes through more and more epochs.

(a)Distorted Loss(GCE) (b) Undistorted loss(CE) (c) Variance Equalizer loss(VE)

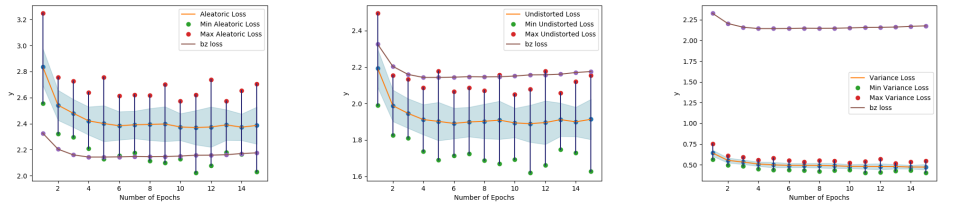

Figure 10: We have shown the variance flow plots over Epochs. This shows role of different type of loss over epoch. from the plot we observed that variance is decreasing as it goes through more and more epochs.

experimented with different ways of placing the dropout layer in the CNN architecture and observe that putting dropout after Max pooling layer works best.

5.2.5. Loss Visualization

We have analyzed the significance of Distorted loss (Gaussian Cross Entropy (GCE) Loss), Undistorted loss (Cross Entropy(CE) Loss) and Variance Equalizer (VE) loss as shown in the Figure 9. It is clear form the figure that all the losses converges as epoch progresses. Variance flow in the various losses is shown in Figure 10. Also we have seen same type of behavior in the variance plot. The variance decreases for all the losses as training progresses.

5.3. Diversity

We used Singular value decomposition (SVD) based evaluation metric to demonstrate diversity across various generated answers in a dialog. Here, we have randomly

Models	R1	R5	R10	MRR	Mean
LF [10]	43.8	74.6	84.0	0.580	5.78
HRE [10]	44.8	74.8	84.3	0.586	5.65
MN [10]	45.5	76.2	85.3	0.596	5.46
HCIAE [24]	48.4	78.7	87.5	0.622	4.81
SF-1 [28]	48.1	78.6	87.5	0.620	4.79
AMEM [54]	48.5	78.6	87.4	0.622	4.85
NMN [12]	**50.9**	80.1	88.8	**0.641**	4.45
ECE (ours)	44.3	76.1	85.9	0.590	5.51
ACE (ours)	49.0	80.5	89.3	0.629	4.32
PDUN (ours)	49.2	**81.0**	**90.5**	0.634	**4.03**

Table 4: In this table, we compare our method performance with state of art results on visual dialog dataset version-0.9[10]. It is shown that our method "PUDN" performs better in all all the scores.except Recall@1. We use same evaluation score as mentioned in the paper. R1 tends, Recall@1, R5 tends Recall@5, R10 tends Recall@10, MRR tends Mean Reciprocal Rank.

selected 400 dialogs. For each dialog, we sampled m number of latent embedding feature using the attentive encoder. Each one is having n-dimensional feature vector. To measure the variance, all the answer embedding features can be concatenated into a feature matrix $A \in R^{m \times n}$. σ_i can be obtained by SVD, $L = U \sum V^T$ of matrix A, where $\sum = diag(\sigma_0, \sigma_1 .... \sigma_{n-1})$; U and V^T are $m \times m$ and $n \times n$ unitary matrices respectively. The overall variance in all dimensions is l_1-norm, $\sigma_o = \sum_{i=0}^{n-1} |\sigma_i|$. A large variance indicates very less correlation among the generated answers, which further implies large diversity among the answers as shown in table 5.

5.4. Comparison with state-of-the-art (SOTA)

The comparison of PDUN method with various state-of-the-art methods for visual dialog dataset v0.9 and v1.0 are provided in table- 4 and table- 6. The first block of the table- 4 and table- 6 consist of the state-of-the-art methods, second block consist of our methods. In table- 4, we compared our proposed 'ECE' model results with baseline results of model 'Late-fusion (LF)' [10] for dataset v0.9. We observe improvement

Model	Diversity(σ_d^2)
VE	06.410
CE	22.231
GCE	27.845
ECE	24.980
ACE	32.120
PDUN	34.350

Table 5: In this table, we provide diversity of answer for various methods on Visual dialog dataset and shows that diversity of "PUDN"method better than individual aleatoric and epistemic method.

Models Mean	R1	R5	R10	MRR	Mean	NDGC
LF [10]	40.9	72.4	82.8	0.55	5.95	0.45
HRE [10]	39.9	70.4	81.5	0.54	6.41	0.45
MN [10]	42.4	74.0	84.3	0.56	5.59	0.47
NMN [12]	**47.5**	78.1	88.8	0.61	4.40	0.54
ECE (ours)	43.6	75.4	85.3	0.58	5.36	0.49
ACE (ours)	47.0	82.4	92.3	0.62	3.81	0.53
PDUN (ours)	47.3	**82.5**	**92.6**	**0.62**	**3.68**	**0.54**

Table 6: In this table, we compare our method performance with state of art results on visual dialog dataset version-1.0[10]. It is shown that our method "PUDN" performs better in all all the scores except Recall@1 and MRR. We use same evaluation score as mentioned in the dataset paper[10]. R1 tends, Recall@1, R5 tends Recall@5, R10 tends Recall@10, MRR tends Mean Reciprocal Rank.

in all scores as compared to base line 'LF' model. The main reason is the baseline mode just concatenates question encoding feature, image encoding feature and history encoding feature to answer the final question. Our proposed 'ECE' model uses probabilistic method to obtain encoding feature with Bayesian CNN and LSTM. We then maximise the diversity using re-parameterization method and minimising the uncertainty. This is used to obtain a improved representation that is used by the decoder stage. In this case we use model uncertainty that is known as epistemic uncertainty. The LF model generalizes poorly due to over-fitting model, overly confident prediction about the input $p(y|\mathbf{x}, \theta)$ which is due to the fact that a single parameter that gener-

ates the data distribution can be easily fooled by adversarial examples. Our 'ECE' model overcome this using probabilistic manner. Our proposed 'ACE' model is similar to 'ECE' model instead of model uncertainty we use data uncertainty i.e. Aleatoric Uncertainty. The main reason behind improve performanced over 'ECE' model is the distorted Gaussian Cross Entropy (GCE) loss. It improves the confidence of prediction by adding noise to the logit input and marginalizes it using 'GCE'. Further we developed 'PUDN' model which combines both epistemic and aleatoric uncertainty. We also compare with Hierarchical Recurrent Encoder (HRE)[10] as another baseline model. Instead of concatenating all the input feature, HRE uses recurrent network to encode history and uses another LSTM model to encode Image, Question and History feature to obtain final encoding feature. We compare with 'MN'[10], which is an attention based memory network as a baseline model with our proposed models. Similarly we compare with other baseline models such as 'HCIAE'[24], 'SF'[28] 'AMEM'[54] and 'NMN'[12]. We see improvement over all of these state-of-the-art baselines by using probabilistic encoders with the help of Bayesian CNN & LSTM with Diversity and Uncertainty module.

Similarly, we compared our proposed 'ECE', 'ACE' and 'PUDN' model results with baseline results as mentioned in table- 6 for version-1.0 of visual dialog dataset. We observe that in terms of @R10 score, we obtain an improvement of around 10% from the baseline & 3.5% from SOTA (NMN [12]) method. In terms of NDGC score 9% from base model & 0.5% from SOTA model and in term of MRR, 7% from the base model & 1% from SOTA (NMN [12]) model using our proposed method.

5.5. Deep Deterministic vs Deep Bayesian Model

We conduct an experiment to compare Bayesian vs non Bayesian model as shown in table-7. We chose late fusion model as a Non Bayesian Deep Model. Then we with the help of drop-out we evaluate our model performance. We choose number of Monte-Carlo sample as 100 to do our output prediction. We see that it performs better than the deterministic model and solves the over-fitting issue. It is also more robust as compared to deterministic model while predicting similar kind of answers. We further extend our model with aleatoric cross entropy loss. We observe that using this, we improve the model performance as compared to the deterministic method.

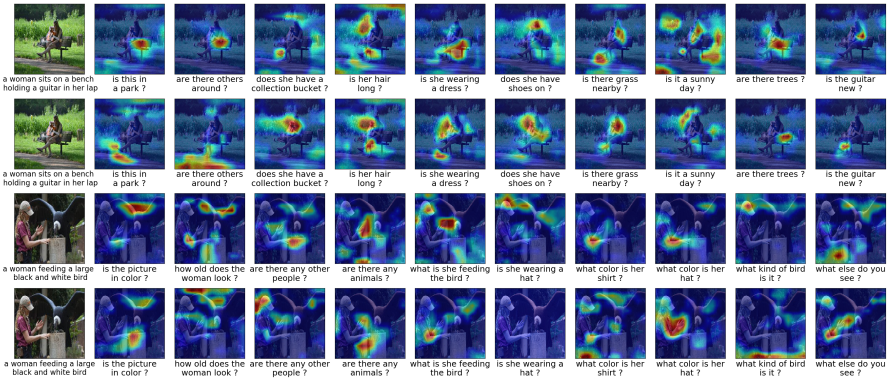

Figure 11: Figure shows the difference between aleatoric dialog results and baseline dialog results. In this figure, the first row refers to Grad-CAM visualization of first example for baseline visual dialog model and second row refers to Grad-CAM visualization of first example for Aleatoric visual dialog model and same scheme is followed for next 2 rows. The first column indicates target Image and corresponding caption and starting from second column is the visualization of rounds of dialog from round 1 to 10.

Method	**R1**	**R5**	**R10**	**MRR**	**Mean**
Late Fusion (Deep Learning)	43.8	74.6	84.0	0.580	5.78
Epistemic Late Fusion (Bayesian Method)	44.0	75.2	85.2	0.593	5.51
Aleatoric Late Fusion (Bayesian Method)	44.3	75.6	85.6	0.585	5.45

Table 7: In this table, we compare performance of deep deterministic model versa Probabilistic model(Bayesian Deep Learning model) on visual dialog dataset. We have show that Bayesian model performs better than non-Bayesian model. We use same evaluation score as mentioned in the paper. R1 tends, Recall@1, R5 tends Recall@5, R10 tends Recall@10, MRR tends Mean Reciprocal Rank.

5.6. Qualitative Result

We provide qualitative results, which easily distinguishes between results of Baseline dialog model with our aleatoric dialog model for two dialog generation example in figure 11. We can clearly see that our proposed method are able to capture uncertainty and minimize it, which further improves dialog results. For example, in the first image, the question is “Is this in a park ?”. The baseline model’s main focus is on the chair, where the uncertainty is very high. But our model explains about field, plant and background image, which provides the extra information about the query that eventually decreases the uncertainty as shown in figure 11. We visualize the certainty activation

map of other two dialogs whose uncertainty score decrease over turns.

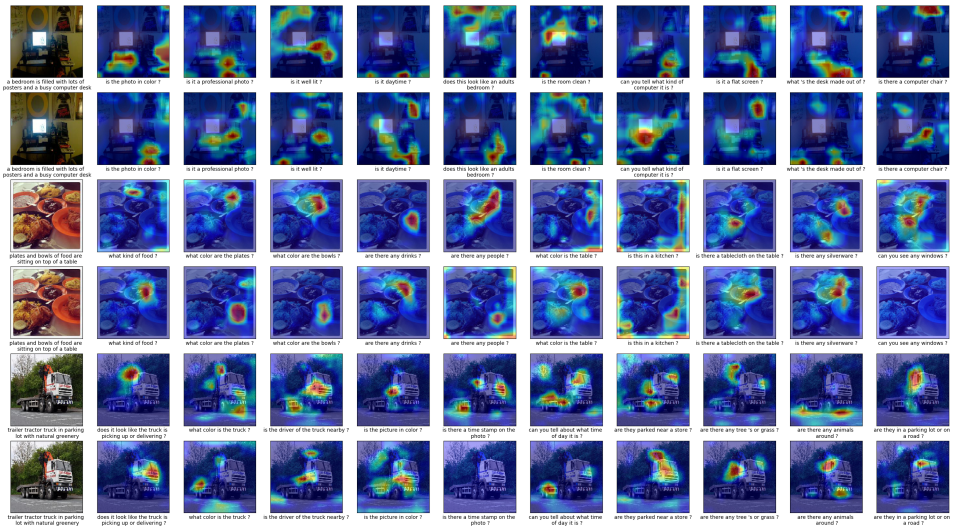

Figure 12: Difference Between Aleatoric dialog results and Baseline dialog results are shown in the figure. In this figure,The first row refer to Grad-CAM visualization of first example for baseline visual dialog model and second row refer to Grad-CAM visualization of first example for aleatoric visual dialog model and so on.. The first column indicates target Image and corresponding caption, second column indicates visualization of dialog round 1, third column refer to visualization of dialog round 2 and so on.

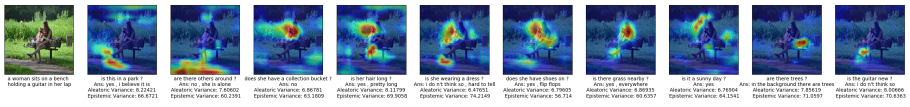

Figure 13: This figure provide aleatoric and epidemic variance and visualize the aleatoric uncertainty using Grad-CAM for a particular Dialog.

We provide qualitative results, which easily distinguishes between results of baseline dialog model with our aleatoric dialog model for three dialog generation example in figure 12. We can clearly see that our proposed method is able to capture uncertainty and minimize it, which further improves dialog results. Also, we have measured epistemic and aleatoric uncertainty and showed how uncertainty decreases as dialog turns in figure 13. We visualize the uncertainty by taking multiple samples and show how does it change as per samples as shown in Figure 14. We also made the GIF version of this visualization with name 'aleatoric_uncertainty_ques_gradcam_100.gif' and other

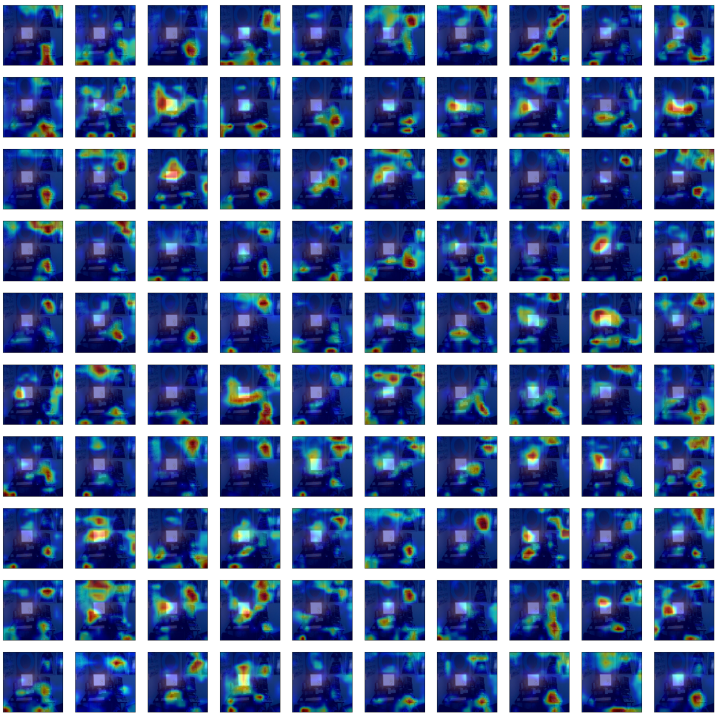

Figure 14: We visualize the multiple outputs from the Bayesian neural network. We took 100 sample from the posterior distribution of dialog model for particular image, particular question. It shows how Grad-CAM is flowing for particular image, particular question.

5.7. Evaluation Protocol

We have followed the evaluation protocol mentioned in [10]. We use a retrieval setting to evaluate the responses at each round in the dialog. Specifically, every question in VisDial is coupled with a list of 100 candidate answers, which the models are asked to sort for evaluation purposes. Models are evaluated on standard retrieval metrics (1) mean rank, (2) recall @k and (3) mean reciprocal rank (MRR) of the human response in the returned sorted list.

[2] `https://delta-lab-iitk.github.io/PDUN/`

5.8. Preprocessing

We truncate captions/questions/answers longer than 24/16/8 words respectively. We then build a vocabulary of words that occur at least 5 times in train, resulting in 8964 words. In our experiments, all 3 Bayesian LSTMs are single layer with 512-dimensional hidden state. For Bayesian CNN we use pretrained VGG-19 [1] with dropout to get the representation of image. We first re-scale the images to 448×448 pixels and take the output of FC7 layer which is 4096-dimensional as image feature. We use the Adam optimizer with a base learning rate of 4e-4.

6. Conclusion

In this paper, we propose a novel probabilistic architecture that we term as the 'Probabilistic diversity and uncertainty network (PDUN)', for solving the problem of visual dialog. The main parts in the proposed architecture are the modules that capture uncertainty and diversity. We capture aleatoric and epistemic uncertainty that provide us with uncertainty estimates and these are further reduced using appropriate loss functions. We have particularly shown that the performance in the visual dialog is improved around 3.5% by the proposed network. Further, the use of the diversity module obtained through a variational autoencoder allows us to generate diverse answers. We validate that indeed the diversity of the proposed network is high as compared to variants of the method. These two contributions enable us to obtain a significantly improved model for solving the challenging visual dialog task. Our model can be used to solve other problems like 'Visual Question Answering' and 'Visual Question Generation' and 'Visual Story Generation'. For these problems we can use our uncertainty and diversity model to generate more natural conversational context. This would be useful for story generation, question and answer generation. In future, we intend to extend our model by improving encoder context with the help of different cues present in the image. We aim to also investigate more useful language models.

7. Acknowledgment

We acknowledge the help provided by our DelTA Lab members and our family who have supported us in our research activity.

8. Reference

References

[1] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[3] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.

[4] Z.-Q. Zhao, P. Zheng, S.-t. Xu, X. Wu, Object detection with deep learning: A review, IEEE transactions on neural networks and learning systems.

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, VQA: Visual Question Answering, in: International Conference on Computer Vision (ICCV), 2015.

[6] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the v in vqa matter: Elevating the role of image understanding in visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6904–6913.

[7] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.

[8] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.

[9] A. Das, S. Kottur, J. M. Moura, S. Lee, D. Batra, Learning cooperative visual dialog agents with deep reinforcement learning, in: IEEE International Conference on Computer Vision (ICCV), 2017.

[10] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, D. Batra, Visual dialog, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[11] B. N. Patro, M. Lunayach, S. Patel, V. P. Namboodiri, U-cam: Visual explanation using uncertainty based class activation maps, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7444–7453.

[12] S. Kottur, J. M. Moura, D. Parikh, D. Batra, M. Rohrbach, Visual coreference resolution in visual dialog using neural module networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 153–169.

[13] Z. Deng, Z. Jiang, R. Lan, W. Huang, X. Luo, Image captioning using densenet network and adaptive attention, Signal Processing: Image Communication (2020) 115836doi:10.1016/j.image.2020.115836.

[14] Y. Wei, L. Wang, H. Cao, M. Shao, C. Wu, Multi-attention generative adversarial network for image captioning, Neurocomputingdoi:10.1016/j.neucom.2019.12.073.

[15] M. Malinowski, M. Fritz, A multi-world approach to question answering about real-world scenes based on uncertain input, in: Advances in Neural Information Processing Systems (NIPS), 2014.

[16] B. Patro, V. P. Namboodiri, Differential attention for visual question answering, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[17] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear attention networks, in: Advances in Neural Information Processing Systems, 2018, pp. 1571–1581.

[18] N. Ruwa, Q. Mao, H. Song, H. Jia, M. Dong, Triple attention network for sentimental visual question answering, Computer Vision and Image Understanding 189 (2019) 102829. doi:10.1016/j.cviu.2019.102829.

[19] W. Li, J. Sun, X. Fang, Visual question answering with attention transfer and a cross-modal gating mechanism, Pattern Recognition Lettersdoi:10.1016/j.patrec.2020.02.031.

[20] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, L. Vanderwende, Generating natural questions about an image, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1802–1813.

[21] U. Jain, Z. Zhang, A. G. Schwing, Creativity: Generating diverse questions using variational autoencoders, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6485–6494.

[22] B. N. Patro, S. Kumar, V. K. Kurmi, V. Namboodiri, Multimodal differential network for visual question generation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 4002–4012.
URL http://aclweb.org/anthology/D18-1434

[23] B. N. Patro, V. K. Kurmi, S. Kumar, V. Namboodiri, Learning semantic sentence embeddings using sequential pair-wise discriminator, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 2715–2729.

[24] J. Lu, A. Kannan, J. Yang, D. Parikh, D. Batra, Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model, in: Advances in Neural Information Processing Systems, 2017, pp. 314–324.

[25] Q. Wu, P. Wang, C. Shen, I. Reid, A. Van Den Hengel, Are you talking to me? reasoned visual dialog generation through adversarial learning, in: Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6106–6115.

[26] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, A. Courville, Guesswhat?! visual object discovery through multi-modal dialogue, in: Proc. of CVPR, 2017.

[27] F. Strub, H. de Vries, J. Mary, B. Piot, A. C. Courville, O. Pietquin, End-to-end optimization of goal-driven and visually grounded dialogue systems, in: IJCAI, 2017.

[28] U. Jain, S. Lazebnik, A. G. Schwing, Two can play this game: Visual dialog with discriminative question generation and answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[29] L. Wang, A. Schwing, S. Lazebnik, Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space, in: Advances in Neural Information Processing Systems, 2017, pp. 5756–5766.

[30] B. Dai, S. Fidler, R. Urtasun, D. Lin, Towards diverse and natural image descriptions via a conditional gan, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 2989–2998.

[31] J. Zhang, Q. Wang, Y. Han, Multi-modal fusion with multi-level attention for visual dialog, Information Processing & Management (2019) 102152.

[32] G. E. Hinton, D. Van Camp, Keeping the neural networks simple by minimizing the description length of the weights, in: Proc. of the Conference on Computational learning theory (COLT), ACM, 1993, pp. 5–13.

[33] C. Blundell, J. Cornebise, K. Kavukcuoglu, D. Wierstra, Weight uncertainty in neural networks (2015) 1613–1622.

[34] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning (ICML), 2016, pp. 1050–1059.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The Journal of Machine Learning Research 15 (1) (2014) 1929–1958.

[36] A. Kendall, V. Badrinarayanan, R. Cipolla, Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding, in: British Machine Vision Conference (BMVC), 2017.

[37] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in: Advances in neural information processing systems, 2017, pp. 5574–5584.

[38] L. Liu, S. Wang, B. Hu, Q. Qiong, J. Wen, D. S. Rosenblum, Learning structures of interval-based bayesian networks in probabilistic generative model for human complex activity recognition, Pattern Recognition 81 (2018) 545–561.

[39] Z. Huang, W. Dong, H. Duan, A probabilistic topic model for clinical risk stratification from electronic health records, Journal of Biomedical Informatics 58 (2015) 28–36.

[40] I. Vulić, W. De Smet, J. Tang, M.-F. Moens, Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications, Information Processing & Management 51 (1) (2015) 111–147.

[41] J. Jiang, X. Li, C. Zhao, Y. Guan, Q. Yu, Learning and inference in knowledge-based probabilistic model for medical diagnosis, Knowledge-Based Systems 138 (2017) 58–68.

[42] C. Bishop, Pattern recognition and machine learning, Journal of Electronic Imaging 16 (2006) 140–155. doi:10.1117/1.2819119.

[43] M. Welling, Y. W. Teh, Bayesian learning via stochastic gradient langevin dynamics, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 681–688.

[44] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in Neural Information Processing Systems, 2017, pp. 6402–6413.

[45] H. Wu, X. Gu, Towards dropout training for convolutional neural networks, Neural Networks 71 (2015) 1–10.

[46] Y. Gal, Z. Ghahramani, A theoretically grounded application of dropout in recurrent neural networks, in: Advances in neural information processing systems, 2016, pp. 1019–1027.

[47] M. Fortunato, C. Blundell, O. Vinyals, Bayesian recurrent neural networks, arXiv preprint arXiv:1704.02798.

[48] D. P. Kingma, M. Welling, Auto-encoding variational bayes, stat 1050 (2014) 10.

[49] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29.

[50] V. K. Kurmi, S. Kumar, V. P. Namboodiri, Attending to discriminative certainty for domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 491–500.

[51] Y. Gal, Uncertainty in deep learning, Ph.D. thesis, University of Cambridge (2016).

[52] K. Dorman, Bayesian neural network blogpost, https://github.com/kyle-dorman/bayesian-neural-network-blogpostAccessed: 2018-08-09.

[53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization., in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[54] P. H. Seo, A. Lehrmann, B. Han, L. Sigal, Visual reference resolution using attention memory for visual dialog, in: Advances in neural information processing systems, 2017, pp. 3719–3729.